The first step was checking and imputing missing values in the dataset. According to the missing value percentage, different technologies were utilized to impute the missing value. If the missing percentage was less than 5%, dropping the rows with the missing value. If the missing value was greater than 5% but smaller than 30% of the data, imputing them by adding new categories or KNN method to predict the missing value using existing features.

The second step was handling the date and time features. There are five features related to date and time in total. I splited them into new features, and each feature represents either year, month, day, hour, minute, or second information.

Next was conducting exploring data analysis(EDA) to help me understand the distribution and relation of the features. Using the library called SweetViz, visualizations of numerical, categorical, and date&time data were conducted separately. The SweetViz report displays the histogram, distribution, association with the target variable, heatmap, and much other statistical information. This step concluded that some features are not associated with the final prediction or highly related to each other and need to be dropped.

Then, I performed data processing, such as missing value imputation, date&time handling, and feature encoding, according to the previous EDA step. For feature encoding, I choose to use both binary encoder and one-hot-encoder. If a categorical feature has too many distinct categories, one-hot-encoding will generate an extremely sparse dataset contains too many features. The too sparse dataset could lead to an overfitting problem. Therefore, for the categorical features with more than 30 distinct features, a binary encoding was conducted. For the rest of the categorical features, the one-hot-encoding strategy was used. I trained many different models with default parameters. In these models, the XGBoost classifier had the best performance. Then I used GridSearchCV in python to tune the parameter. Using the best parameter, I predicted the fraud-nonfraud in the test set.